Lab #5

By the beginning of the next lab (Feb. 24), send what you have to [afodor@uncc.edu](mailto:afodor@uncc.edu)

Make sure the text “Lab #5” is in the subject line…

(1) Download the file http://afodor.github.io/classes/stats2015/longitdunalRNASeqData.zip

(2) Read the counts table ( “nc101\_scaff\_dataCounts.txt “ into R). For example:

setwd("C:\\classes\\AdvancedStats\_Spring2016")

myT <- read.table("nc101\_scaff\_dataCounts.txt",sep="\t",header=TRUE,row.names=1)

numCols <- ncol(myT)

myColClasses <- c("character", rep("numeric", numCols))

myTAsNum <-read.table("nc101\_scaff\_dataCounts.txt",sep="\t",header=TRUE,colClasses=myColClasses)

(myTAsNum tells R that these are numbers and not factors, which it can sometimes get confused about!)

(3) Normalize the spreadsheet using a simple normalization scheme (that is divide each sample by the number of samples in the sample).

(For example:

myTNorm <- myTAsNum

for ( i in 2:ncol(myTAsNum))

{

colSum = sum(myTNorm[,i])

myTNorm[,i] = myTNorm[,i]/colSum

}

(4) In myTNorm, columns 2:4 are the 2 week time point, 5:7 are the 12 week timepoint, 8:12 are the 20 week timepoint). As a preview for next week (and since our “snow” day messed up my schedule), to run a t-test comparing week 2 and week 20 for the first gene, we could use:

t.test( myTNorm[ 1,2:4 ], myTNorm[ 1, 8:12] )$p.value

For all three comparisons (2 weeks vs. 12 weeks; 2 weeks vs. 20 weeks; 12 weeks vs. 20 weeks), run t-tests over all the genes. Plot histograms of the resulting p-values. Are any of the of the p-values uniform? Given the three histograms of p-values, which two timepoints do you think would have the most significant differences?

(5) Given a vector of pValues, one can adjust for multiple hypothesis testing with (for example):

p.adjust( pValuesFromTTest[!is.nan(pValuesFromTTest)], method="BH")

At a 10% false discovery rate, how many genes are significant for the three comparisons under the t-test?